# TASK-AWARE SEPARATION FOR THE DCASE 2020 TASK 4 SOUND EVENT DETECTION AND SEPARATION CHALLENGE

*Samuele Cornell*[1*], *Michel Olvera*[2*] *Manuel Pariente*[2*], *Giovanni Pepe*[1*], *Emanuele Principi*[1*], *Leonardo Gabrielli*[1], *Stefano Squartini*[1]

[1] Università Politecnica delle Marche, Dept. Information Engineering, Ancona, Italy, {s.cornell;g.pepe}@pm.univpm.it, {e.principi;l.gabrielli;s.squartini}@univpm.it
[2] INRIA Nancy Grand-Est, Dept. Information and Communication Sciences and Technologies, France {manuel.pariente;michel.olvera}@inria.fr

## ABSTRACT

Source Separation is often used as a pre-processing step in many signal-processing tasks. In this work we propose a novel approach for combined Source Separation and Sound Event Detection in which a Source Separation algorithm is used to enhance the Sound Even -Detection back-end performance. In particular, we present a permutation-invariant training scheme for optimizing the Source Separation system directly with the back-end Sound Event Detection objective without requiring joint training or fine-tuning of the two systems. We show that such an approach has significant advantages over the more standard approach of training the Source Separation system separately using only a Source Separation based objective such as Scale-Invariant Signal-To-Distortion Ratio. On the 2020 Detection and Classification of Acoustic Scenes and Events Task 4 Challenge our proposed approach is able to outperform the baseline source separation system by more than one percent in event-based macro $F_1$ score on the development set with significantly less computational requirements.

***Index Terms—*** Source Separation, End-to-End, Sound Event Detection, Time Domain, Joint Training

## 1. INTRODUCTION

Source separation aims at extracting from an acoustic mixture its underlying acoustic components. In recent years, Deep-Neural-Networks (DNN) based methods have made significant progress towards this goal especially in the speech processing field [1, 2] but also regarding the more general problem of separating arbitrary sounds [3]. In the speech processing field, source separation has a wide variety of applications such as hearing aid devices [4], Automatic Speech Recognition [5, 6, 7] (ASR) and diarization [8]. At the same time, arbitrary sound separation also has potential applications in hearing devices and in creative applications such as video editing.

A recent work [9] has demonstrated that sound classification can be used to improve the performance of an arbitrary sound separation algorithm but the inverse problem of whether a source separation algorithm can be used to improve sound classification and more in general Sound Event Detection (SED) is still a matter of research. In fact, answering such a question is one of the goals of the 2020 Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 Task 4 Sound Event Detection (SED) and Separation challenge. To motivate participants to explore this direction and develop a combined source separation and SED system the challenge organizers have provided a baseline sound separation model and a dataset, the Free Universal Sound Separation (FUSS) [10], suitable for training such arbitrary sound separation algorithm.

In this paper, we propose a new technique for combined source separation and SED and show it can improve the back-end SED performance on DCASE 2020 Task 4 development set. In our approach, instead of optimizing separately the separation system with a loss function focused only on source separation, we leverage the pre-trained back-end SED system to optimize the separation system also with the SED objective. In this way, we assure that separation will be helpful to the back-end SED system and will not introduce a mismatch which can deteriorate the performance. This approach, is inspired by previous techniques aimed at improving ASR performance with Source Separation [11, 12, 7] and Speech Enhancement [13, 14] by using either joint training or ASR related losses in order to ensure that separation or enhancement will be beneficial to the ASR task. We compare our proposed method with the aforementioned DCASE 2020 Task 4 combined separation and SED baseline and we show that the proposed approach outperforms it with significantly less parameters.

This work is organized as follows: in Section 2 we briefly describe the datasets available in the context of the DCASE2020 Task 4 Sound Event Detection and Separation challenge; in Section 3 we describe the challenge combined source separation and sound event detection baseline and, following, our proposed approach. In Section 4 we show the results obtained, perform an in-depth ablation study to assess the validity of our work, and discuss the strengths and weaknesses of the proposed method. Finally, in Section 5 we draw conclusions and outline possible future research directions.

## 2. DCASE 2020 TASK 4 CHALLENGE DATASETS

The DCASE 2020 Task 4 challenge is focused primarily on Sound Event Detection in real-world domestic environments with weakly-annotated data, unlabeled data, and only a very small corpus of strongly annotated, out of domain, synthetic data.

Three datasets are available for training the SED systems: DESED [15], SINS [16] and TUT [17]. The SINS and TUT Acous-

*Equal contributions

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

tic scenes 2017 datasets offer only background noise while the DESED [4] dataset offers, for training, weakly labeled and unlabeled real soundscapes, and isolated synthetic events with strong labels. It is a dataset primarily made for SED systems training and evaluation. The synthetic portion has background noise derived from SINS and because oracle foregrounds events are available it could also be used for training a source separation algorithm. Regarding validation and evaluation, instead, DESED has only real soundscapes with strong labels in order to assess algorithms generalization to unseen, real-world scenarios. It features a total of 10 different sounds events classes for training a SED system: Speech, Dog, Cat, Alarm bell/ringing, Dishes, Frying, Blender, Running water, Vacuum cleaner, Electric shaver/toothbrush.

As the challenge is also focused on Source Separation, an additional, out of domain dataset is made available for training Source Separation systems: the Free Universal Sound Separation (FUSS) Dataset [10]. Being originally aimed at arbitrary sound separation [3], it offers isolated events and noise backgrounds but no SED annotations. Moreover, it does come from a completely different domain than DESED as it does not include the same classes but also arbitrary audio events which the SED system should ideally ignore.

## 3. COMBINED SOURCE SEPARATION AND SOUND EVENT DETECTION

Hereafter we present and discuss our proposed technique for tackling SED using Source Separation as a pre-processing step in the context of the DCASE 2020 Task 4 Sound Event Detection and Separation Challenge.

In our experiments for combined separation and SED we used the released pre-trained SED baseline system together with our separation system[1]. This allows to directly compare with results reported by the baseline combined source separation and SED system. The baseline SED system is derived from [18] and is based on a hybrid Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architecture trained using both weakly, strong and unlabeled data by using a Mean-Teacher strategy [19] for Semi-Supervised training. It uses 128 log-Mel features in input taken with a 2048 samples window and 128 hop size. The network predictions are smoothed with a median filter with a 7 frames length.

### 3.1. DCASE 2020 Combined Source Separation and Sound Event Detection Baseline System

The most straightforward approach to perform combined SED and source separation is to train the Source Separation system and SED separately using different objectives. This is also the approach adopted by the baseline source separation system.

The official source separation baseline is derived from [10] and is trained on a synthetic dataset comprised of FUSS as well as synthetic examples from DESED. This baseline model is optimized with an End-to-End (E2E) waveform-based Scale-Invariant Signal-to-Distortion-Ratio (SI-SDR) [20] objective to remove the background noise from the mixtures. Thus it performs denoising rather than full foregrounds separation from the mixtures. The whole approach is illustrated in Figure 1. The network architecture is based on TDCN++ [3, 9] and follows the analysis/masking/synthesis scheme where a DNN is used to estimate a mask in a transformed domain for each source. The masking is performed in Short-Time-Fourier-Transform (STFT) magnitude spectra domain.

[1] Available at `github.com/turpaultn/dcase20_task4`.

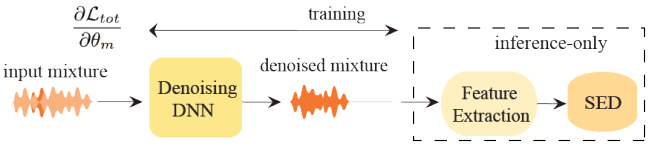


Figure 1: DCASE 2020 Task 4 baseline separation system approach.

A potential drawback is that this approach does not guarantee that the denoised mixtures will be more suitable for SED. In fact, the denoising process could lead to mixtures whose distribution is significantly different from the one of noisy mixtures. Because the baseline SED model is trained on noisy mixtures directly, the denoising process can potentially introduce a mismatch. This issue is common in speech processing applications where, for example, Speech-Enhancement denoising is used as a pre-processing step for ASR [13, 14, 21]. This can explain the modest performance improvement given by the baseline separation model.

Moreover, this modest improvement is only reached using ensembling: the SED model is fed both the denoised and noisy audio features and the predictions are averaged. This allows for improvement over the baseline SED-only system but at the expense of raising considerably the computational requirements.

### 3.2. Task-Aware Separation

On the contrary, we depart significantly from this common procedure and propose what we call Task-Aware separation training to address the aforementioned domain mismatch problem that arises when separation/denoising is performed on a system trained on noisy mixtures.

For this reason, it is often preferable to jointly train from scratch the separation system and SED system in an E2E fashion or jointly fine-tune the two pre-trained systems. In this way, the separation system is guaranteed to help the back-end SED system task as it is trained with the SED objective. This method has been shown to be highly effective for joint source separation and ASR [11, 12, 7] even surpassing monolithic, multi-talker, Permutation Invariant Training ASR techniques such as [22, 23].

Another, more common, procedure to guarantee that the back-end SED system is matched to the separated/denoised pre-processed data is to re-train or fine-tune the SED system on this data. However, this is a rather expensive procedure as it requires separate training of separation system and then separate fine-tuning or re-training of the SED algorithm.

Hereafter we present another approach which we call Task-Aware separation, which allows to train a separation system using a pre-trained SED back-end with the SED objective, thus avoiding the domain mismatch problem. A significant advantage of our procedure over joint training is that potentially a robust back-end, pre-trained on a vast amount of data, for which oracle targets for separation are not available, can be directly used.

Our approach is illustrated in Figure 2. As the combined system ultimate goal is SED, we perform Deep Neural Network (DNN) mask-based separation directly on Mel-spectrograms. The separated features are then fed to the pre-trained SED system after applying logarithm and scaling. We then use both the predictions of the SED as well as its internal activations to train the mask-estimation DNN network. In the whole process, the back-end SED

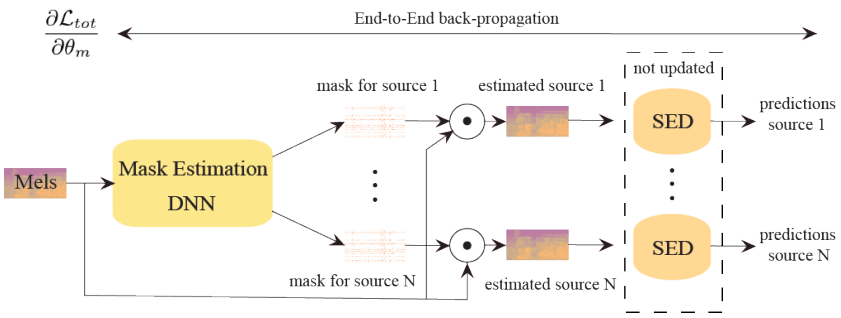


Figure 2: Task-aware Separation for Sound Event Detection.

model is not updated, but gradients are back-propagated through it in order to update the mask-estimation network.

We used the same data used for the challenge baseline SED training, comprised of synthetic (for which oracle foregrounds are available), weakly, and unlabeled examples. For synthetic examples we used dynamic mixing of sources and backgrounds: we construct synthetic training examples at training time by randomly sampling from one to five random DESED foregrounds and one background from SINS. We then apply reverberation to each source independently by using FUSS Room Impulse Responses (RIRs) and mix the foregrounds and background. The level for each foreground is randomly sampled between -35 dB and 0 dB while the background is constrained to be at max 5 dB over the foreground which has minimum level. FUSS was not employed for training. In fact, we found that adding FUSS did bring a slight performance loss due to the fact that it features a completely different domain than DESED. Moreover, we found that some DESED foreground classes were considered as backgrounds in FUSS.

Permutation Invariant Training (PIT) [24, 25] and Mean Teacher [19] are used to train the mask-estimation DNN. Because the sound class are well defined, it can be argued that the use of PIT is not necessary. However, in our preliminary experiments, we found that the separation stack trained with a non-PIT loss led to overfitting the weakly and synthetic examples very quickly. By using PIT this problem is mitigated, as the separation stack is not anymore required to explicitly recognize the sound classes. In addition, the use of PIT opens to future developments for different applications such as speaker separation.

We denote with $\mathbf{f} = [f_j(t)]_{j=1\dots J}^{t=1\dots N}$ and $\widehat{\mathbf{f}} = [\widehat{f}_j(t)]_{j=1\dots J}^{t=1\dots N}$ the matrices of true and estimated targets, where $J$ and $N$ are respectively the maximum number of sources and the length of the estimated and true targets. The PIT-equivalent loss of a generic loss function $\mathcal{L}$ can be written as:

$$\mathcal{L}_{\text{PIT}} = \min_{\sigma \in \mathcal{F}_{\mathcal{J}}} \mathcal{L}(\widehat{\mathbf{f}}_\sigma, \mathbf{f}), \tag{1}$$

where $\widehat{\mathbf{f}}_\sigma = [\widehat{f}_{\sigma(j)}(t)]_{j=1\dots J}^{t=1\dots N}$ is a permutation of $\mathbf{f}$ by $\sigma \in \mathcal{F}_{\mathcal{J}}$, with $\mathcal{F}_{\mathcal{J}}$ the set of permutations of $[1, ..., J]$. The whole procedure consists in computing the loss $\mathcal{L}$ for all possible permutations of the targets and finding the permutation $\sigma$ for which the loss is minimized. We used the implementation of PIT available in Asteroid Source Separation Toolkit [26].

Several different losses depending on what labels are available for the current example are used to train the mask-estimation DNN:

- **Strongly Labelled**: for DESED synthetic data, for which foregrounds features $\mathbf{f} = [f_j(t)]_{j=1\dots J}^{t=1\dots N}$ are available we compute PIT Mean-Squared Error loss $\mathcal{L}_{MSE}$ and find the optimal permutation $\sigma_{opt}$ for the estimated separated features $\widehat{\mathbf{f}} = [\widehat{f}_j(t)]_{j=1\dots J}^{t=1\dots N}$:

$$\mathcal{L}_{\text{MSE}} = \min_{\sigma \in \mathcal{F}_{\mathcal{J}}} MSE(\widehat{\mathbf{f}}_\sigma, \mathbf{f}). \tag{2}$$

The estimated foregrounds features are then re-ordered according to $\sigma_{opt}$ and fed to the SED model for computing Deep Feature Loss (DFL) [27, 28]. DFL $\mathcal{L}_{DFL}$ is computed between each SED internal activations obtained with re-ordered estimated foregrounds $SED(\widehat{\mathbf{f}_{\sigma_{opt}}})$ and those obtained with oracle foregrounds $SED(\mathbf{f})$:

$$\mathcal{L}_{\text{DFL}} = \sum_{m=1}^{M} \left\| SED(\widehat{\mathbf{f}_{\sigma_{opt}}})^m - SED(\mathbf{f})^m \right\|_1, \tag{3}$$

where the sum is taken over all M layers of the SED back-end and $SED(\mathbf{f})^m$ denotes the activations of the m-th layer when the SED model is fed the feature matrix $\mathbf{f}$. As in [28], $L_1$ norm is used for computing DFL. The total loss for strongly labelled examples is the sum of the two terms:

$$\mathcal{L}_{\text{strong}} = \mathcal{L}_{DFL} + \mathcal{L}_{\text{MSE}} \tag{4}$$

- **Weakly Labelled**: for weakly labelled data no oracle foregrounds are available, thus we train the separation model as in E2E Neural Diarization [29] to minimize PIT binary cross entropy (BCE) between weak predictions of SED model when it is fed the estimated foregrounds features $\widehat{\mathbf{w}}_\sigma = SED(\widehat{\mathbf{f}}_\sigma)_{weak}$ and the weak labels $\mathbf{w}_{weak}$:

$$\mathcal{L}_{\text{weak}} = \min_{\sigma \in \mathcal{W}_{\mathcal{J}}} BCE(\widehat{\mathbf{w}}_\sigma, \mathbf{w}_{weak}). \tag{5}$$

- **Mean-teacher consistency**: to leverage also unlabeled data, we use the Mean Teacher Semi-Supervised loss for the mask-estimation network and enforce SED weak and strong predictions consistency between the values obtained with a student separation model $\mathbf{S}(\mathbf{f}; \boldsymbol{\theta}_s)$ and an exponential moving average mean teacher separation model $\mathbf{T}(\mathbf{f}; \boldsymbol{\theta}_t)$ using permutation invariant MSE loss $\mathcal{L}_{teach}$ between the separated features of the two models:

$$\mathcal{L}_{\text{teach}} = \min_{\sigma \in \mathcal{F}_{\mathcal{J}}} MSE(SED(\mathbf{S}(\mathbf{f}_\sigma)), SED(\mathbf{T}(\mathbf{f}))). \tag{6}$$

The total loss $\mathcal{L}_{\text{tot}}$ used to train the mask-estimation DNN is then the sum of aforementioned strong, weak, and mean-teacher losses.

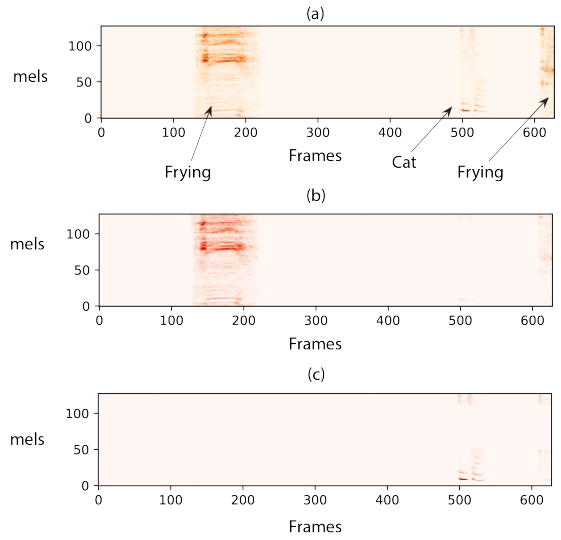


Figure 3: Output of separation stack: (a) input mixture, (b) first estimated source (c) second estimated source.

## 4. RESULTS AND DISCUSSION

In our experiments, mask-estimation was accomplished with a reduced version of the separator network from Conv-TasNet [2], as implemented in Asteroid, with 5 blocks ($X = 5$) and 3 repeats ($R = 3$), 64 bottleneck channels, 128 depth-wise convolution channels and sigmoid mask for a total of 3M trainable parameters. The whole system was trained to separate a maximum of 5 different sound event classes. We used a batch size of 32 examples with 12 synthetic examples, 12 weakly labeled examples, and 8 unlabeled examples, respectively. Adam [30] was used for optimization. We used a warm-up exponential learning rate schedule for the first 10 epochs with the learning rate increasing from 0 to 0.001 as in [18].

In the following, we report results results obtained on DCASE 2020 Task 4 development set. We report results only in terms of SED event-based macro $F_1$ score as oracle sources required for calculating Source Separation metrics such as SI-SDR are only available for the training set.

### 4.1. DCASE 2020 Task 4 Results

In Table 1 we report the results of the proposed separation system trained with the Task-Aware separation objective (Proposed). We also report the performance attained by the SED challenge baseline back-end system on its own without pre-processing (SED-only Baseline), the performance of the combined separation and sound event detection baseline with (SEP+SED Baseline) and without averaging of predictions between noisy and denoised mixtures (SEP+SED Baseline no avg). In addition, we report the total number of parameters for each model with the back-end SED model included.

As can be seen, the combined separation and SED baseline system fails to improve the SED back-end performance when no ensembling is performed, while ensembling produces only a moderate performance improvement. On the other hand, the proposed method offers more than 2% improvement over the plain SED-only baseline with no ensembling and a significantly smaller separation model. In Figure 3 we show the output of the proposed separation system (in Mel domain) for a mixture where three different sound events are present belonging to two distinct classes: Frying and Cat. It can be seen that the events are effectively separated by the system, but that also distortion is introduced. However, E2E training guarantees that this distortion does not harm SED performance.

| Method | Event macro $F_1$ score | Parameters |
|---|---|---|
| SED-only Baseline | 34.8 | 1M |
| SEP+SED Baseline | 35.6 | 10M |
| SEP+SED Baseline no avg | 33.4 | 10M |
| Proposed | 37.0 | 4M |

Table 1: Performance of combined separation and SED systems on DCASE 2020 Task 4 development set.

### 4.2. Ablation Study

In Table 2 we report results for our separation system when different loss functions are used. As a baseline, we report the system trained with only the PIT MSE loss $\mathcal{L}_{\text{MSE}}$ (Equation 2) computed with the synthetic examples (strong-PIT), thus with only a source-separation based objective. We report again the performance of the SED back-end alone and of the combined separation and SED baseline.

It can be seen that the system trained with only signal-based separation loss slightly degrades the performance compared to the SED-only system. This is due to the separation system quickly overfitting the synthetic data, leading to poor generalization to real-world examples. If the weak examples loss (+weak) $\mathcal{L}_{\text{weak}}$ (Equation 5) is added, the separation system is able to improve over the SED-only back-end system. The addition of Deep Feature Loss (Equation 3) brings another substantial improvement while the addition of the semi-supervised Mean-Teacher loss (Equation 6) adds only marginal improvement.

| Method | Event macro $F_1$ score |
|---|---|
| SED-only | 34.8 |
| SEP+SED Baseline | 35.6 |
| strong-MSE | 34.5 |
| +weak | 35.4 |
| +weak+DFL | 36.7 |
| +weak+DFL+teach | 37 |

Table 2: Ablation study for proposed technique (development set).

## 5. CONCLUSIONS

In this paper, we have proposed a novel training scheme for combined Source Separation and Sound Event Detection. In our method, a source separation system is trained in an End-to-End fashion with a pre-trained SED system that is not updated. For this purpose, we devised a combination of permutation-invariant training objectives, both signal-based and SED-based. We call such an approach Task-Aware separation as the separation system is optimized directly with the back-end task objective. We compared our approach with the state-of-the-art combined separation and Sound Event Detection DCASE 2020 Task 4 baseline and we showed that such a method is able to reach superior performance on DCASE 2020 Task 4 development set while having significantly less parameters. In future works, we will expand this End-to-End approach to other back-end tasks such as Automatic Speech Recognition.

## 6. REFERENCES


[1] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*. IEEE, 2016, pp. 31–35.

[2] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[3] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 175–179.

[4] K. Reindl, Y. Zheng, and W. Kellermann, "Speech enhancement for binaural hearing aids based on blind source separation," in *ISCCSP*. IEEE, 2010, pp. 1–6.

[5] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," *Interspeech*, 2018.

[6] T. Menne, I. Sklyar, R. Schlüter, and H. Ney, "Analysis of deep clustering as preprocessing for automatic speech recognition of sparsely overlapping speech," *arXiv preprint arXiv:1905.03500*, 2019.

[7] T. von Neumann, K. Kinoshita, L. Drude, C. Boeddeker, M. Delcroix, T. Nakatani, and R. Haeb-Umbach, "End-to-end training of time domain audio separation and recognition," in *ICASSP*. IEEE, 2020, pp. 7004–7008.

[8] T. von Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *ICASSP*. IEEE, 2019, pp. 91–95.

[9] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *ICASSP*. IEEE, 2020, pp. 96–100.

[10] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *in preparation*, 2020.

[11] A. Narayanan and D. Wang, "Improving robustness of deep neural network acoustic models via speech separation and joint adaptive training," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 92–101, 2014.

[12] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *ICASSP*. IEEE, 2018, pp. 4819–4823.

[13] S. Watanabe, T. Hori, J. Le Roux, and J. R. Hershey, "Student-teacher network learning with enhanced features," in *ICASSP*. IEEE, 2017, pp. 5275–5279.

[14] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *ICASSP*. IEEE, 2018, pp. 5609–5613.

[15] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[16] G. Dekkers, S. Lauwereins, B. Thoen, M. W. Adhana, H. Brouckxon, B. Van den Bergh, T. van Waterschoot, B. Vanrumste, M. Verhelst, and P. Karsmakers, "The sins database for detection of daily activities in a home environment using an acoustic sensor network," *Detection and Classification of Acoustic Scenes and Events 2017*, 2017.

[17] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *EUSIPCO*, Budapest, Hungary, 2016.

[18] L. Delphin-Poulat and C. Plapous, "Mean teacher with data agumentation for dcase 2019 task 4," in *DCASE 2019 Tech Report*, 2009.

[19] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[20] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP*. IEEE, 2019, pp. 626–630.

[21] P. Wang, K. Tan, *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[22] H. Seki, T. Hori, S. Watanabe, J. L. Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *arXiv preprint arXiv:1805.05826*, 2018.

[23] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *ICASSP*. IEEE, 2019, pp. 6256–6260.

[24] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *ICASSP*. IEEE, 2017, pp. 241–245.

[25] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, 2017.

[26] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, *et al.*, "Asteroid: the pytorch-based audio source separation toolkit for researchers," *arXiv preprint arXiv:2005.04132*, 2020.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[28] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[29] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-End Neural Speaker Diarization with Permutation-Free Objectives," in *Interspeech*, 2019, pp. 4300–4304.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.